02450

Introduction to Machine Learning and Data Mining

Project 2

20.04.21



Avi Raj
s210252

Piotr Samir Saffrani
s210364

Pranjal Garg
s210242

Table 1: Students contribution to the report

Section	Avi Raj	Piotr Saffarini	Pranjal Garg
Introduction	40%	35%	25%
Regression Part a	15%	42.5%	42.5%
Regression Part b	15%	42.5%	42.5%
Classification	45%	20%	35%
Discussion	40%	30%	30%
Exam Questions	40%	30%	30%

Contents

1 Introduction

The regression problem that was defined in last report was to predict adiposity based on other attributes in the South African Heart Disease dataset. Through this regression task, a relationship between adiposity and the other attributes was established. Adiposity level of an individual is the measurement of body-fat percentage. Therefore it is a good parameter to for an individual to judge their fitness level. This study can be useful for medical experts and individuals alike to measure the impact of a person's fitness levels on the other attributes.

Using the correlation plot for adiposity, a subset of most relevant attributes was selected to predict adiposity. Using the selected attributes, a Linear Regression model was used for prediction and find out the best regularization parameter in Regression part a. Later on in Regression part b, different regression models mainly baseline(mean-based), Linear Regression with regularization parameter and Artificial Neural Networks with number of hidden units as a parameter.

In the classification task, the presence or absence of CHD is predicted; meaning whether the person has coronary heart disease or not. All other attributes have been used to predict CHD class label as the prediction whether someone would have a CHD event could help in saving their life and if all the information available, it would make a better prediction. three models for classification have been used mainly baseline based on largest class label, logistic regression and Decision Trees.

2 Regression part a

2.1 Variable Prediction, Regression Goal and Feature Transformation

Adiposity is the feature that is being predicted in this report. All the 10 attributes that are present in our dataset are shown in Table 2. In the project 1 report, a strong correlation was found for adipostiy with age and obesity. After further analysis of the correlation of adiposity, it can be seen in the correlation plot that adiposity has a fairly relevant correlation with ldl, sbp, tobacco and CHD as well.

Obesity has the strongest correletion with adiposity. But obesity by its sheer definition is very similar to adiposity and carries the same information in the real world. Hence it would be better idea to omit obesity as one of the features for regression as it wont provide any new valuable information about their relationship and would affect the analysis on other attributes.

Hence age,ldl, sbp, tobacco and CHD are used for the regression analysis.

Before running the regression models, the data was standardized so that all attributes except binary attributes have mean 0 and standard deviation 1. There was no other transformation necessary for regression. One out of K was encoding was tried for CHD but omitted in the end as the feature is binary and was leading to an unnecessary increase of dimensionality without any added benefits.

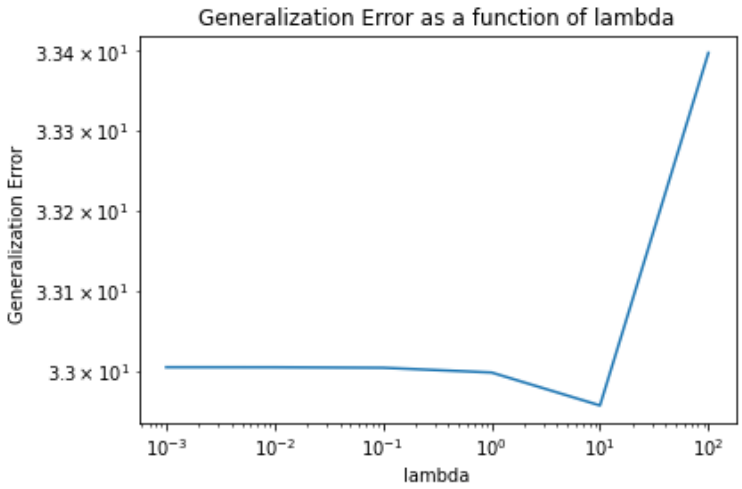
Table 2: Data features

sbp	systolic blood pressure (mmHg)
tobacco	cumulative lifetime tobacco usage (kg)
ldl	low density lipoprotein cholesterol (mmol/L)
adiposity	The body adiposity is calculated value of body fat (%)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	Body Mass Index (kg/m^2)
alcohol	alcohol consumption (l)
age	age at onset (yrs)
chd	response, coronary heart disease (Positive, Negative)

2.2 Introducing λ parameter

The model chosen to estimate adiposity value was logistic regression with λ parameter (also called ridge regression [1]). Lambda was introduced to penalize the model for incorrect approximation in the data set. It adds some bias to the approximation just to make it less vulnerable to real data samples. Lambda is iteratively adjusted via 10 fold splits in inner fold of the K-fold.

Figure 1: Generalization error on different values of λ

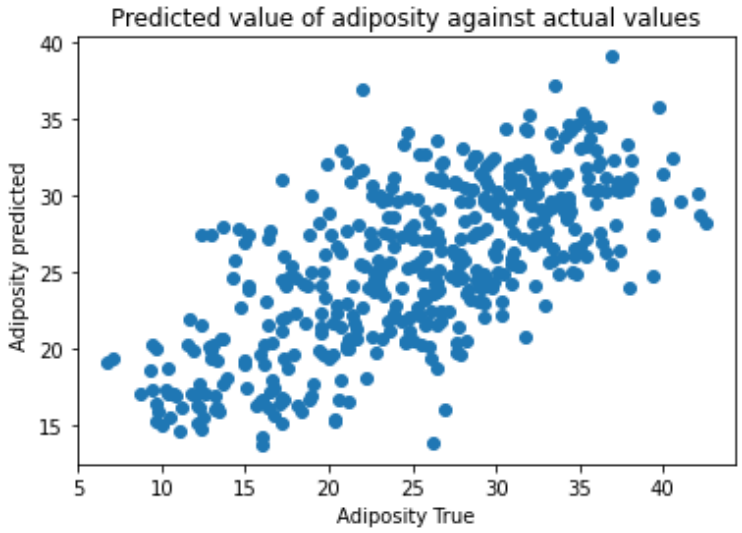


2.3 Data prediction on the best model

Below on figure 2 is the plot comparing actual vs. predicted estimation of adiposity. The plot was computed on the value of λ with the lowest generalization error. Ideally, the points should create a straight line with 45 °angle to X axis.

The predictions are not perfect, but points form thick diagonal line which is a good sign. Including obesity as an input data would increase the accuracy of the model, but it was not used as explained before in Section 1.1 Variable Prediction, Regression Goal and Feature Transformation.

Figure 2: Predicted values of adiposity against actual



3 Regression part b

3.1 Two level cross validation

Double layer cross validation was performed on the data to compare 3 regression calculation methods: Artificial Neural Networks (ANN), Linear regression and baseline which was mean of the training set. Values used to choose the optimal lambda in ridge regression and h (hidden layers) in ANN were as followed:

$\lambda\{10^{-3}, 10^{-2}, ..., 10^{2}\}$

h {1, 2, 3, 4, 5}

The main computational power was taken by the ANN, significantly increasing the time spent on running the code. So in the report only 5 h values were taken into consideration. As calculations of λ were quick, there was not any reason against performing optimization of this parameter on broader range, but after couple of runs the range of 5 values from 0.001 to 100 were chosen.

3.2 Comparing regression models

The computed values of parameters and generalization errors of all regression models mentioned are described in the Table 3 below. From the table it is clear that ANN has the lowest generalization error on the test data of all the methods. As the difference is not marginal, it makes a strong argument for ANN, making it the model that would perform best in comparison to all other tested models in any autonomous system. The optimal values of h and λ were consistent for every fold, as h = 3 and $\lambda = 10$. The best lambda value was the same as chosen in Regression part a.

Table 3: Two-level cross-validation table used to compare the three models

Outer fold	**ANN**		**Linear regression**		**Baseline**
i	h_i	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	5.92	10	30.62	55.68
2	3	8.49	10	30.85	70.11
3	4	6.85	10	30.45	46.45
4	3	7.17	10	26.57	64.10
5	3	7.61	10	33.61	60.94
6	3	6.89	10	29.41	57.70
7	3	9.29	10	37.34	69.84
8	3	7.26	10	29.93	43.77
9	3	10.86	10	43.35	69.84
10	3	8.57	10	37.38	73.66

3.3 Comparing regression p values

After computing all models and their best parameters, next step is to compare their performance in the most objective way possible. Grading the models by looking only at generalization errors is not reproducible and may lead to wrong conclusions. This is why setup II [3] for statistical evaluation was carried out. This method takes the results from different training sets which is crucial for estimating how the compared models would perform outside the researched case scenario. The results are presented below in Table 4.

To help reject or accept the null hypothesis which states that generalization errors of two compared models are equal, p-value [2] was introduced. The null hypothesis was rejected to all of the comparisons, because p-values were far below 0.05. Obtaining

negative confidence interval values qualified the first model of the two compared as better suited for our particular case. It shows linear regression is better than baseline and ANN was better than linear regression. In conclusion, the most optimal model for this dataset is the Artificial Neural Networks model with three hidden units.

Table 4: Setup II for evaluating performance difference

compared models	**p value**	**Con**
Linear Regression vs. Baseline	0.000002540	(-34.39, -22.12)
ANN vs. Baseline	0.000000025	(-60.07, -46.53)
ANN vs. Linear Regression	0.000000008	(-27.85, -22.24)

4 Classification

4.1 Goal and type of classification.

For the classification problem CHD is predicted. All the remaining attributes available are used for predicting the presence of coronary heart disease in an individual. Our classification problem is a binary classification problem as CHD is either present(1) or absent(0)

To analyse any data it is crucial that the data is in the best possible format. For this reason all the attributes of the type,'Ordinal','Ratio','Interval' have been standardized. This means the mean of all these attributes is converted to zero and the standard deviation converted to one. For the attributes like 'famhist', one out of K encoding was tried but ultimately abandoned as it just increased the dimentionality without adding any information. Also due to multicolinearity issues led to dummy variable trap. Hence one out of K encoding was not done on family history.

4.2 Selection of optimal parameter for complexity control

For classification, a baseline based on largest class, logistic regression and Decision Tree Classification was used. For the Decision Tree Classification, 'Gini Impurity' was used as an impurity measure. For controlling the complexity of the Decision Tree model, the option was between 'impurity calculation method' and 'minimum splits'. The former does not provide a lot of complexity as there are two options for it: 'Entropy' or 'Gini impurity'. Hence, minimum number of splits was used which has a wide range from a minimum of two splits to numerous splits.

For baseline classification a model , mode was used as the prediction and for logistic regression, ridge regression was used with λ parameter [1]). Lambda was introduced to penalize the model for incorrect approximation in the data set. It adds some bias to the approximation just to make it less vulnerable to real data samples.

4.3 Cross Validation

4.4 Statistical Evaluations

Similar to the statistical evaluation performed in regression, setup II was used for statistical evaluation and the results are shown in the Table 6.

To help reject or accept the null hypothesis for classification, p-value [2] was used similar to regression. The p-value for Decision Tree vs Baseline is greater than 0.05 , so the null hypothesis is accepted so both Decision Tree and baseline perform in a similar way. The confidence interval also reflects this conclusion as one value is positive and one value is negative. The null hypothesis was rejected for all the other comparisons i.e logistic regression vs baseline and logistic regression vs Decision Trees, because p-values were far below 0.05. Negative confidence interval means the the first model is better than the second and positive

Table 5: Two-level cross-validation table used to compare the three models.

Outer fold	Decision Trees		Logistic Regression		Baseline
i	x_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	7	0.31914894	10	0.20573693	0.5106383
2	5	0.31914894	10	0.17762067	0.29787234
3	2	0.41304348	10	0.17911999	0.30434783
4	4	0.41304348	10	0.179902	0.34782609
5	7	0.45652174	100	0.21719759	0.34782609
6	7	0.41304348	10	0.22506947	0.41304348
7	6	0.30434783	10	0.15125963	0.30434783
8	3	0.30434783	100	0.15932862	0.32608696
9	3	0.39130435	100	0.17534433	0.39130435
10	11	0.30434783	10	0.16291929	0.2173913

confidence interval means the the second model is better. Hence the table shows that Logistic regression is better than baseline and Decision Trees and Decision Trees and baseline are very similar. So it can concluded that Logistic Regression is the best model for the classification problem.

Table 6: Setup II for evaluating performance difference

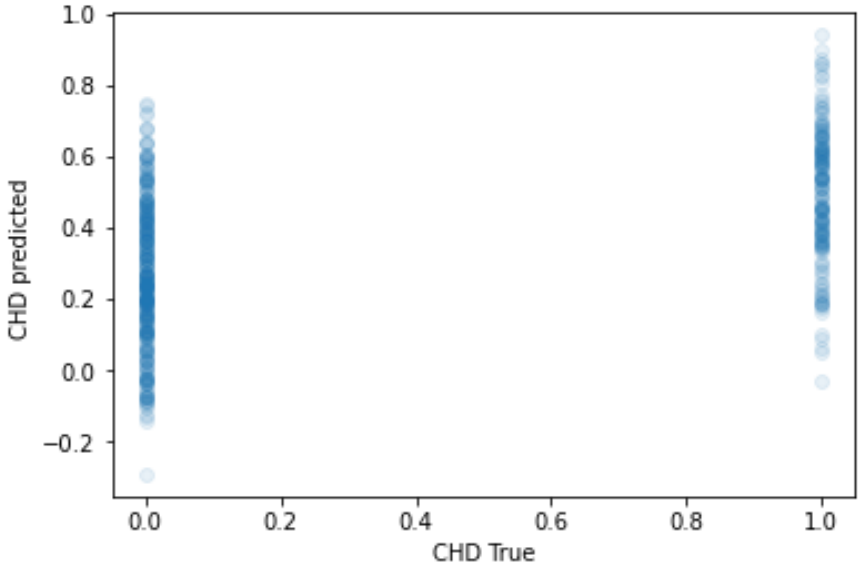
Compared Models	**p value**	**Confidence Interval**
Decision Tree vs Baseline	0.539739769	(-0.037, 0.054)
Logistic Regression vs Baseline	0.000028803	(-0.210, -0.115)
Decision Tree vs. Logistic Regression	0.000000707	(0.133, 0.2095)

4.5 Logistic Regression Model

Logistic regression is especially good at classifying categorical data. The algorithm works by fitting a sigmoid curve into the graph between the dependent and independent variables. It then segregates the values into different classes depending on where the independent variable lies on the curve.Several such curves are tried on the dataset and it is found when one which has the least error is selected where λ =10 . The graph below shows the values of predicted CHD vs actual values. Looking at the graph, it can be said that for true values the predicted value of CHD is generally higher than for false values.

Family history, age and tobacco, are found to be most relevant attributes in the mentioned order when it comes to classification. In regression, age, ldl, and sbp are the most relevant But since regression is for adiposity values with different features and classification is for CHD with a different set of features, the results are not surprising.

Figure 3: Predicted values of CHD against actual



5 Discussion

5.1 Summary & conclusions

Throughout this project we have gained a deeper insight into the internal workings of various classification and regression algorithms was developed.

In the regression part, the adiposity of an individual was predicted to a strong degree. On evaluating the data with three models, the conclusion was that Artificial Neural Networks or ANN is the most suited model of the three to make predictions on the chosen attributes.

For the classification section of the project predicted the presence of Coronary Heart Disease or CHD in a person was predicted. All the nine attributes available were used to get the highest amount of accuracy. For method 2, 'Decision Trees' was used, which was also reasonably accurate. It was observed that Logistic Regression had the most accurate results.

Through this project the importance of machine learning in the field of medical science and healthcare can be realised. If individuals and medical experts can obtain insights of possible ailments in the future, it could significantly improve the help and support of medical experts to the patients. In the first project report, we asked whether it is possible for an individual to assess their lifestyle and make changes to reduce their chances of CHD and now we can conclude that lifestyle factors can be used to predict CHD and hence changes in them would improve an individual chances of not contracting CHD.

5.2 Previous Study Analysis

Two separate studies were analysed. The first study had conducted regression analysis on the dataset [4], whereas the second one had made their predictions using classification algorithms [5].

The first study performs several regression models on the dataset mainly 'CART','baseline','Boosting','Linear Regression'. According to that study boosting model has the best accuracy. The study found that alcohol consumption does not have any large effect on the predictions of the models and hence might not be a important attribute for predicting CHD. On the other hand, 'Age' and 'Tobacco' were the most important attributes in all the models.'SBP' was deemed useful only by the boosting model.

The second study was more focused on detecting CHD than finding the leading causes. The study used all the attributes and performed several classification modellings to find the model that best suits the dataset. The algorithms used were,'Logistic regression','Random Forest','Naive Bayes','Support vector Machines','Gradient Boost'. According to the study,random forest shows the best results. To calculate this the study calculated the following parameters for every study,'Accuracy','Confusion Matrix','ROC Curves', and 'F1 score'.

6 Exam Questions

6.0.1 Question1

Answer is option C.

Prediction C is right because if you start trying to plot the ROC curve using the candidate predictions in Figure 2 of the question, then as you move from right to left of all the plots in figure 2(i.e starting with high threshold values, moving from $\hat{y}$ = 1 to $\hat{y}$ = 0 for threshold values), then you should find a positive value first as TPR = 0.25(1 out of 4 true positive values is detected as positive and 3 are classified as false negatives) and FPR=0(no true negative values is classified as positive). And as observations are made along the ROC curve in figure 1, and moving right to left on figure 2,Prediction C is found to be the right plot, that would result in the ROC curve in figure 1

6.0.2 Question 2

Answer is option C.

For X_7= 2, the following was obtained-
Left Split y_1 = 0,y_2 = 1,y_3 = 0,y_4=0
Right split: y_1 = 37, y_2 = 20, y_3 = 33, y_4=34

I(r) = 1- 37/135 =98/135

I(Left) = 1-1/1 = 0

I(Right)= 1-37/134= 97/134

N(v_right)= 134, N(v_left)= 1, N(r)= 135

N(v_right)/N(r) = 134/135

purity gain= 98/135 -((134/135)*(97/134)) $\approx$ 1/135 = 0.0074

6.0.3 Question 3

Answer is option C.

In an ANN, the number of parameters is simply the sum of the product of the numbers of nodes in connected layers. The input layer has 7 nodes, the hidden layer has 10 nodes and the output layer has 4 nodes

Total number of parameters= (7*10) + (10*4) = 110

6.0.4 Question 4

Answer is option D.

It's obvious from Figure 4. If you use the equation, b1=> -0.76, and look at the area that is the complement of this equation, it covers the entire area of congestion 2 and some area of congestion 1 and then further if you look at b2=>0.03, then it divides the area further into congestion 1 and 2. So A is b1=> -0.76, and B is b2=> 0.03

Now if you look at the true side of the equation b1=¿ -0.76 and further divide it using the equation b1¿=-0.16 then it separates congestion level 4 from congestion level 3 and 1. And further in the area of congestion level 3 and 1, if equation b2¿=0.01 is used then it divides the remaining area in congestion level 3 and congestion level 1 separately. So C is b1=¿ -0.76 and D is b2¿=0.01

Hence A is b1=> -0.76, B is b2=>0.03, C is b1=> -0.76 and D is b2>=0.01

6.0.5 Question 5

Answer is option C. The total time to calculate the table is 3570 s.

K1=5, K2=4

Number of times ANN is trained and tested= K1*((K2* number of hidden units) + 1)= 5*((4* 5) + 1) = 105

Number of times regression is trained and tested= K1*((K2* no of distinct lambdas) + 1) = 5*((4* 5) + 1) = 105

Total time spent= 105 * (training ANN time + Testing ANN time) +105 * (training LR time + Testing LR time) = 3570 ms

6.0.6 Question 6

Option B is correct After computing this equation over every possible answer: P(y = 4 | $\hat{y}$)= 1+ $\sum_{k'=1}^{3} e^{\hat{y}_k}$ Every option had possibility on the level of 1e-6 besides B which was about 0.7. Code in python below:

```
b1 = 0.7, b2 = 3.8

y_k = np.array([1, b1, b2])

w1 = np.array([1.2, -2.1, 3.2])

w2 = np.array([1.2, -1.7, 2.9])

w3 = np.array([1.3, -1.1, 2.2])

y1 = np.matmul(y_k.transpose(), w1)

y2 = np.matmul(y_k.transpose(), w2)

y3 = np.matmul(y_k.transpose(), w3)

prob = 1/(1+np.exp(y1)+ np.exp(y2)+ np.exp(y3))

print(prob)
```

References

[1] Ridge regression explained
`https://machinelearningmastery.com/ridge-regression-with-python/`

[2] A. Baum: Course 02402 Introduction to Statistics
`Lecture 5: Hypothesis testing`

[3] Tue Herlau, Mikkel N. Schmidt and Morten Mørup: Introduction to Machine Learning and Data Mining
`Lecture notes, Spring 2020, version 1.0 pages 191-210`

[4] First previous study: Prediction of Coronary Heart Disease by learning from retrospective study
`http://srisai85.github.io/CHD/heartattack.html`

[5] Second previous study: Machine Learning Algorithms for Coronary Heart Disease Prediction
`https://antoniocastiglione-9550.medium.com/machine-learning-algorithms-for-coronary-heart-disease-prediction-ec25f4d7ee42`